

Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse

Elon Musk is famous for his futuristic gambles, but Silicon Valley's latest rush to embrace artificial intelligence scares him. And he thinks you should be frightened too. Inside his efforts to influence the rapidly advancing field and its proponents, and to save humanity from machine-learning overlords.

by [Maureen Dowd](#)

March 26, 2017 5:00 pm

<https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x>

I. Running Amok

It was just a friendly little argument about the fate of humanity. Demis Hassabis, a leading creator of advanced artificial intelligence, was chatting with [Elon Musk](#), a leading doomsayer, about the perils of artificial intelligence.

They are two of the most consequential and intriguing men in Silicon Valley who don't live there. Hassabis, a co-founder of the mysterious London laboratory DeepMind, had come to Musk's SpaceX rocket factory, outside Los Angeles, a few years ago. They were in the canteen, talking, as a massive rocket part traversed overhead. Musk explained that his ultimate goal at SpaceX was the most important project in the world: interplanetary colonization.

Hassabis replied that, in fact, *he* was working on the most important project in the world: developing artificial super-intelligence. Musk countered that this was one reason we needed to colonize Mars—so that we'll have a bolt-hole if A.I. goes rogue and turns on humanity. Amused, Hassabis said that A.I. would simply follow humans to Mars.

This did nothing to soothe Musk's anxieties (even though he says there are scenarios where A.I. wouldn't follow).

An unassuming but competitive 40-year-old, Hassabis is regarded as the Merlin who will likely help conjure our A.I. children. The field of A.I. is rapidly developing but still far from the powerful, self-evolving software that haunts Musk. Facebook uses A.I. for targeted advertising, photo tagging, and curated news feeds. Microsoft and Apple use A.I. to power their digital assistants, Cortana and Siri. Google's search engine from the beginning has been dependent on A.I. All of these small advances are part of the chase to eventually create flexible, self-teaching A.I. that will mirror human learning.

WITHOUT OVERSIGHT, MUSK BELIEVES, A.I. COULD BE AN EXISTENTIAL THREAT: “WE ARE SUMMONING THE DEMON.”

Some in Silicon Valley were intrigued to learn that Hassabis, a skilled chess player and former video-game designer, once came up with a game called *Evil Genius*, featuring a malevolent scientist who creates a doomsday device to achieve world domination. [Peter Thiel](#), the billionaire venture capitalist and [Donald Trump](#) adviser who co-founded PayPal with Musk and others—and who in December helped gather skeptical Silicon Valley titans, including Musk, for [a meeting with the president-elect](#)—told me a story about an investor in DeepMind who joked as he left a meeting that he ought to shoot Hassabis on the spot, because it was the last chance to save the human race.

Elon Musk began warning about the possibility of A.I. running amok three years ago. It probably hadn't eased his mind when one of Hassabis's partners in DeepMind, Shane Legg, stated flatly, “I think human extinction will probably occur, and technology will likely play a part in this.”

Before DeepMind was gobbled up by Google, in 2014, as part of its A.I. shopping spree, Musk had been an investor in the company. He told me that his involvement was not about a return on his money but rather to keep a wary eye on the arc of A.I.: “It gave me more visibility into the rate at which things were improving, and I think they're really improving at an accelerating rate, far faster than people realize. Mostly because in everyday life you don't see robots walking around. Maybe your Roomba or something. But Roombas aren't going to take over the world.”

In a startling public reproach to his friends and fellow techies, Musk warned that they could be creating the means of their own destruction. He told Bloomberg's Ashlee Vance, the author of the biography *Elon Musk*, that he was afraid that his friend [Larry Page](#), a co-founder of Google and now the C.E.O. of its parent company, Alphabet, could have perfectly good intentions but still “produce something evil by accident”—including, possibly, “a fleet of artificial intelligence-enhanced robots capable of destroying mankind.”

At the World Government Summit in Dubai, in February, Musk again cued the scary organ music, evoking the plots of classic horror stories when he noted that “sometimes what will happen is a scientist will get so engrossed in their work that they don't really realize the ramifications of what they're doing.” He said that the way to escape human obsolescence, in the end, may be by “having some sort of merger of biological intelligence and machine intelligence.” This Vulcan mind-meld could involve something called a neural lace—an injectable mesh that would literally hardwire your brain to communicate directly with computers. “We're already cyborgs,” Musk told me in February. “Your phone and your computer are extensions of you, but the interface is through finger movements or speech, which are very slow.” With a neural lace inside your skull you would flash data from your brain, wirelessly, to your digital devices or to virtually unlimited computing power in the cloud. “For a meaningful partial-brain interface, I think we're roughly four or five years away.”

Musk's alarming views on the dangers of A.I. first went viral after he spoke at M.I.T. in 2014—speculating (pre-Trump) that A.I. was probably humanity's “biggest existential threat.” He added that he was increasingly inclined to think there should be some national or international

regulatory oversight—anathema to Silicon Valley—“to make sure that we don’t do something very foolish.” He went on: “With artificial intelligence, we are summoning the demon. You know all those stories where there’s the guy with the pentagram and the holy water and he’s like, yeah, he’s sure he can control the demon? Doesn’t work out.” Some A.I. engineers found Musk’s theatricality so absurdly amusing that they began echoing it. When they would return to the lab after a break, they’d say, “O.K., let’s get back to work summoning.”

Musk wasn’t laughing. “Elon’s crusade” (as one of his friends and fellow tech big shots calls it) against unfettered A.I. had begun.

II. “I Am the Alpha”

Elon Musk smiled when I mentioned to him that he comes across as something of an Ayn Rand-ian hero. “I have heard that before,” he said in his slight South African accent. “She obviously has a fairly extreme set of views, but she has some good points in there.”

But Ayn Rand would do some re-writes on Elon Musk. She would make his eyes gray and his face more gaunt. She would refashion his public demeanor to be less droll, and she would not countenance his goofy giggle. She would certainly get rid of all his nonsense about the “collective” good. She would find great material in the 45-year-old’s complicated personal life: his first wife, the fantasy writer Justine Musk, and their five sons (one set of twins, one of triplets), and his much younger second wife, the British actress Talulah Riley, who played the boring Bennet sister in the Keira Knightley version of *Pride & Prejudice*. Riley and Musk were married, divorced, and then re-married. They are now divorced again. Last fall, Musk tweeted that Talulah “does a great job playing a deadly sexbot” on HBO’s *Westworld*, adding a smiley-face emoticon. It’s hard for mere mortal women to maintain a relationship with someone as insanely obsessed with work as Musk.

“How much time does a woman want a week?” he asked Ashlee Vance. “Maybe ten hours? That’s kind of the minimum?”

Mostly, Rand would savor Musk, a hyper-logical, risk-loving industrialist. He enjoys costume parties, wing-walking, and Japanese steampunk extravaganzas. Robert Downey Jr. used Musk as a model for Iron Man. Marc Mathieu, the chief marketing officer of Samsung USA, who has gone fly-fishing in Iceland with Musk, calls him “a cross between Steve Jobs and Jules Verne.” As they danced at their wedding reception, Justine later recalled, Musk informed her, “I am the alpha in this relationship.”

In a tech universe full of skinny guys in hoodies—whipping up bots that will chat with you and apps that can study a photo of a dog and tell you what breed it is—Musk is a throwback to Henry Ford and Hank Rearden. In *Atlas Shrugged*, Rearden gives his wife a bracelet made from the first batch of his revolutionary metal, as though it were made of diamonds. Musk has a chunk of one of his rockets mounted on the wall of his Bel Air house, like a work of art.

Musk shoots for the moon—literally. He launches cost-efficient rockets into space and hopes to eventually inhabit the Red Planet. In February he announced plans to send two space tourists on

a flight around the moon as early as next year. He creates sleek batteries that could lead to a world powered by cheap solar energy. He forges gleaming steel into sensuous Tesla electric cars with such elegant lines that even the nitpicking [Steve Jobs](#) would have been hard-pressed to find fault. He wants to save time as well as humanity: he dreamed up the Hyperloop, an electromagnetic bullet train in a tube, which may one day whoosh travelers between L.A. and San Francisco at 700 miles per hour. When Musk visited secretary of defense Ashton Carter last summer, he mischievously tweeted that he was at the Pentagon to talk about designing a Tony Stark-style “flying metal suit.” Sitting in traffic in L.A. in December, getting bored and frustrated, he tweeted about creating the Boring Company to dig tunnels under the city to rescue the populace from “soul-destroying traffic.” By January, according to *Bloomberg Businessweek*, Musk had assigned a senior SpaceX engineer to oversee the plan and had started digging his first test hole. His sometimes quixotic efforts to save the world have inspired a parody twitter account, “Bored Elon Musk,” where a faux Musk spouts off wacky ideas such as “Oxford commas as a service” and “bunches of bananas genetically engineered” so that the bananas ripen one at a time.

Of course, big dreamers have big stumbles. Some SpaceX rockets have blown up, and last May a driver was killed in a self-driving Tesla whose sensors failed to notice the tractor-trailer crossing its path. (An investigation by the National Highway Traffic Safety Administration found that Tesla’s Autopilot system was not to blame.)

Musk is stoic about setbacks but all too conscious of nightmare scenarios. His views reflect a dictum from *Atlas Shrugged*: “Man has the power to act as his own destroyer—and that is the way he has acted through most of his history.” As he told me, “we are the first species capable of self-annihilation.”

Here’s the nagging thought you can’t escape as you drive around from glass box to glass box in Silicon Valley: the Lords of the Cloud love to yammer about turning the world into a better place as they churn out new algorithms, apps, and inventions that, it is claimed, will make our lives easier, healthier, funnier, closer, cooler, longer, and kinder to the planet. And yet there’s a creepy feeling underneath it all, a sense that we’re the mice in their experiments, that they regard us humans as Betamaxes or eight-tracks, old technology that will soon be discarded so that they can get on to enjoying their sleek new world. Many people there have accepted this future: we’ll live to be 150 years old, but we’ll have machine overlords.

VIDEO: [Elon Musk Multitasks Better Than You](#)

Maybe we already have overlords. As Musk slyly told Recode’s annual Code Conference last year in Rancho Palos Verdes, California, [we could already be playthings in a simulated-reality world](#) run by an advanced civilization. Reportedly, two Silicon Valley billionaires are working on an algorithm to break us out of the Matrix.

Among the engineers lured by the sweetness of solving the next problem, the prevailing attitude is that empires fall, societies change, and we are marching toward the inevitable phase ahead. They argue not about “whether” but rather about “how close” we are to replicating, and

improving on, ourselves. Sam Altman, the 31-year-old president of Y Combinator, the Valley's top start-up accelerator, believes humanity is on the brink of such invention.

"The hard part of standing on an exponential curve is: when you look backwards, it looks flat, and when you look forward, it looks vertical," he told me. "And it's very hard to calibrate how much you are moving because it always looks the same."

You'd think that anytime Musk, Stephen Hawking, and Bill Gates are all raising the same warning about A.I.—as all of them are—it would be a 10-alarm fire. But, for a long time, the fog of fatalism over the Bay Area was thick. Musk's crusade was viewed as Sisyphean at best and Luddite at worst. The paradox is this: Many tech oligarchs see everything they are doing to help us, and all their benevolent manifestos, as streetlamps on the road to a future where, as Steve Wozniak says, humans are the family pets.

But Musk is not going gently. He plans on fighting this with every fiber of his carbon-based being. Musk and Altman have founded OpenAI, a billion-dollar nonprofit company, to work for safer artificial intelligence. I sat down with the two men when their new venture had only a handful of young engineers and a makeshift office, an apartment in San Francisco's Mission District that belongs to Greg Brockman, OpenAI's 28-year-old co-founder and chief technology officer. When I went back recently, to talk with Brockman and Ilya Sutskever, the company's 30-year-old research director (and also a co-founder), OpenAI had moved into an airy office nearby with a robot, the usual complement of snacks, and 50 full-time employees. (Another 10 to 30 are on the way.)

Altman, in gray T-shirt and jeans, is all wiry, pale intensity. Musk's fervor is masked by his diffident manner and rosy countenance. His eyes are green or blue, depending on the light, and his lips are plum red. He has an aura of command while retaining a trace of the gawky, lonely South African teenager who immigrated to Canada by himself at the age of 17.

In Silicon Valley, a lunchtime meeting does not necessarily involve that mundane fuel known as food. Younger coders are too absorbed in algorithms to linger over meals. Some just chug Soylent. Older ones are so obsessed with immortality that sometimes they're just washing down health pills with almond milk.

At first blush, OpenAI seemed like a bantamweight vanity project, a bunch of brainy kids in a walkup apartment taking on the multi-billion-dollar efforts at Google, Facebook, and other companies which employ the world's leading A.I. experts. But then, playing a well-heeled David to Goliath is Musk's specialty, and he always does it with style—and some useful sensationalism.

Let others in Silicon Valley focus on their I.P.O. price and ridding San Francisco of what they regard as its unsightly homeless population. Musk has larger aims, like ending global warming and dying on Mars (just not, he says, on impact).

Musk began to see man's fate in the galaxy as his personal obligation three decades ago, when as a teenager he had a full-blown existential crisis. Musk told me that *The Hitchhiker's Guide to the*

Galaxy, by Douglas Adams, was a turning point for him. The book is about aliens destroying the earth to make way for a hyperspace highway and features Marvin the Paranoid Android and a supercomputer designed to answer all the mysteries of the universe. (Musk slipped at least one reference to the book into the software of the Tesla Model S.) As a teenager, Vance writes in his biography, Musk formulated a mission statement for himself: “The only thing that makes sense to do is strive for greater collective enlightenment.”

OpenAI got under way with a vague mandate—which isn’t surprising, given that people in the field are still arguing over what form A.I. will take, what it will be able to do, and what can be done about it. So far, public policy on A.I. is strangely undetermined and software is largely unregulated. The Federal Aviation Administration oversees drones, the Securities and Exchange Commission oversees automated financial trading, and the Department of Transportation has begun to oversee self-driving cars.

Musk believes that it is better to try to get super-A.I. first and distribute the technology to the world than to allow the algorithms to be concealed and concentrated in the hands of tech or government elites—even when the tech elites happen to be his own friends, people such as Google founders Larry Page and Sergey Brin. “I’ve had many conversations with Larry about A.I. and robotics—many, many,” Musk told me. “And some of them have gotten quite heated. You know, I think it’s not just Larry, but there are many futurists who feel a certain inevitability or fatalism about robots, where we’d have some sort of peripheral role. The phrase used is ‘We are the biological boot-loader for digital super-intelligence.’ ” (A boot loader is the small program that launches the operating system when you first turn on your computer.) “Matter can’t organize itself into a chip,” Musk explained. “But it can organize itself into a biological entity that gets increasingly sophisticated and ultimately can create the chip.”

Musk has no intention of being a boot loader. Page and Brin see themselves as forces for good, but Musk says the issue goes far beyond the motivations of a handful of Silicon Valley executives.

“It’s great when the emperor is Marcus Aurelius,” he says. “It’s not so great when the emperor is Caligula.”

III. The Golden Calf

After the so-called A.I. winter—the broad, commercial failure in the late 80s of an early A.I. technology that wasn’t up to snuff—artificial intelligence got a reputation as snake oil. Now it’s the hot thing again in this go-go era in the Valley. Greg Brockman, of OpenAI, believes the next decade will be all about A.I., with everyone throwing money at the small number of “wizards” who know the A.I. “incantations.” Guys who got rich writing code to solve banal problems like how to pay a stranger for stuff online now contemplate a vertiginous world where they are the creators of a new reality and perhaps a new species.

Microsoft’s Jaron Lanier, the dreadlocked computer scientist known as the father of virtual reality, gave me his view as to why the digerati find the “science-fiction fantasy” of A.I. so tantalizing: “It’s saying, ‘Oh, you digital techy people, you’re like gods; you’re creating life;

you're transforming reality.' There's a tremendous narcissism in it that we're the people who can do it. No one else. The Pope can't do it. The president can't do it. No one else can do it. We are the masters of it The software we're building is our immortality." This kind of God-like ambition isn't new, he adds. "I read about it once in a story about a golden calf." He shook his head. "Don't get high on your own supply, you know?"

Google has gobbled up almost every interesting robotics and machine-learning company over the last few years. It bought DeepMind for \$650 million, reportedly beating out Facebook, and built the Google Brain team to work on A.I. It hired Geoffrey Hinton, a British pioneer in artificial neural networks; and Ray Kurzweil, the eccentric futurist who has predicted that we are only 28 years away from the Rapture-like "Singularity"—the moment when the spiraling capabilities of self-improving artificial super-intelligence will far exceed human intelligence, and human beings will merge with A.I. to create the "god-like" hybrid beings of the future.

It's in Larry Page's blood and Google's DNA to believe that A.I. is the company's inevitable destiny—think of that destiny as you will. ("If evil A.I. lights up," Ashlee Vance told me, "it will light up first at Google.") If Google could get computers to master search when search was the most important problem in the world, then presumably it can get computers to do everything else. In March of last year, Silicon Valley gulped when a fabled South Korean player of the world's most complex board game, Go, was beaten in Seoul by DeepMind's AlphaGo. Hassabis, who has said he is running an Apollo program for A.I., called it a "historic moment" and admitted that even he was surprised it happened so quickly. "I've always hoped that A.I. could help us discover completely new ideas in complex scientific domains," Hassabis told me in February. "This might be one of the first glimpses of that kind of creativity." More recently, AlphaGo played 60 games online against top Go players in China, Japan, and Korea—and emerged with a record of 60--0. In January, in another shock to the system, an A.I. program showed that it could bluff. Libratus, built by two Carnegie Mellon researchers, was able to crush top poker players at Texas Hold 'Em.

Peter Thiel told me about a friend of his who says that the only reason people tolerate Silicon Valley is that no one there seems to be having any sex or any fun. But there are reports of sex robots on the way that come with apps that can control their moods and even have a pulse. The Valley is skittish when it comes to female sex robots—an obsession in Japan—because of its notoriously male-dominated culture and its much-publicized issues with sexual harassment and discrimination. But when I asked Musk about this, he replied matter-of-factly, "Sex robots? I think those are quite likely."

VIDEO: Silicon Valley's Buffer Zones

Whether sincere or a shrewd P.R. move, Hassabis made it a condition of the Google acquisition that Google and DeepMind establish a joint A.I. ethics board. At the time, three years ago, forming an ethics board was seen as a precocious move, as if to imply that Hassabis was on the verge of achieving true A.I. Now, not so much. Last June, a researcher at DeepMind co-authored a paper outlining a way to design a "big red button" that could be used as a kill switch to stop A.I. from inflicting harm.

Google executives say Larry Page's view on A.I. is shaped by his frustration about how many systems are sub-optimal—from systems that book trips to systems that price crops. He believes that A.I. will improve people's lives and has said that, when human needs are more easily met, people will "have more time with their family or to pursue their own interests." Especially when a robot throws them out of work.

Musk is a friend of Page's. He attended Page's wedding and sometimes stays at his house when he's in the San Francisco area. "It's not worth having a house for one or two nights a week," the 99th-richest man in the world explained to me. At times, Musk has expressed concern that Page may be naïve about how A.I. could play out. If Page is inclined toward the philosophy that machines are only as good or bad as the people creating them, Musk firmly disagrees. Some at Google—perhaps annoyed that Musk is, in essence, pointing a finger at them for rushing ahead willy-nilly—dismiss his dystopic take as a cinematic cliché. Eric Schmidt, the executive chairman of Google's parent company, put it this way: "Robots are invented. Countries arm them. An evil dictator turns the robots on humans, and all humans will be killed. Sounds like a movie to me."

Some in Silicon Valley argue that Musk is interested less in saving the world than in buffing his brand, and that he is exploiting a deeply rooted conflict: the one between man and machine, and our fear that the creation will turn against us. They gripe that his epic good-versus-evil story line is about luring talent at discount rates and incubating his own A.I. software for cars and rockets. It's certainly true that the Bay Area has always had a healthy respect for making a buck. As Sam Spade said in *The Maltese Falcon*, "Most things in San Francisco can be bought, or taken."

Musk is without doubt a dazzling salesman. Who better than a guardian of human welfare to sell you your new, self-driving Tesla? Andrew Ng—the chief scientist at Baidu, known as China's Google—based in Sunnyvale, California, writes off Musk's Manichaeian throwdown as "marketing genius." "At the height of the recession, he persuaded the U.S. government to help him build an electric sports car," Ng recalled, incredulous. The Stanford professor is married to a robotics expert, issued a robot-themed engagement announcement, and keeps a "Trust the Robot" black jacket hanging on the back of his chair. He thinks people who worry about A.I. going rogue are distracted by "phantoms," and regards getting alarmed now as akin to worrying about overpopulation on Mars before we populate it. "And I think it's fascinating," he said about Musk in particular, "that in a rather short period of time he's inserted himself into the conversation on A.I. I think he sees accurately that A.I. is going to create tremendous amounts of value."

Although he once called Musk a "sci-fi version of P. T. Barnum," Ashlee Vance thinks that Musk's concern about A.I. is genuine, even if what he can actually do about it is unclear. "His wife, Talulah, told me they had late-night conversations about A.I. at home," Vance noted. "Elon is brutally logical. The way he tackles everything is like moving chess pieces around. When he plays this scenario out in his head, it doesn't end well for people."

Eliezer Yudkowsky, a co-founder of the Machine Intelligence Research Institute, in Berkeley, agrees: "He's Elon-freaking-Musk. He doesn't need to touch the third rail of the artificial-intelligence controversy if he wants to be sexy. He can just talk about Mars colonization."

Some sniff that Musk is not truly part of the whiteboard culture and that his scary scenarios miss the fact that we are living in a world where it's hard to get your printer to work. Others chalk up OpenAI, in part, to a case of FOMO: Musk sees his friend Page building new-wave software in a hot field and craves a competing army of coders. As Vance sees it, "Elon wants all the toys that Larry has. They're like these two superpowers. They're friends, but there's a lot of tension in their relationship." A rivalry of this kind might be best summed up by a line from the vainglorious head of the fictional tech behemoth Hooli, on HBO's *Silicon Valley*: "I don't want to live in a world where someone else makes the world a better place better than we do."

Musk's disagreement with Page over the potential dangers of A.I. "did affect our friendship for a while," Musk says, "but that has since passed. We are on good terms these days."

Musk never had as close a personal connection with 32-year-old Mark Zuckerberg, who has become an unlikely lifestyle guru, setting a new challenge for himself every year. These have included wearing a tie every day, reading a book every two weeks, learning Mandarin, and eating meat only from animals he killed with his own hands. In 2016, it was A.I.'s turn.

Zuckerberg has moved his A.I. experts to desks near his own. Three weeks after Musk and Altman announced their venture to make the world safe from malicious A.I., Zuckerberg posted on Facebook that his project for the year was building a helpful A.I. to assist him in managing his home—everything from recognizing his friends and letting them inside to keeping an eye on the nursery. "You can think of it kind of like Jarvis in Iron Man," he wrote.

One Facebooker cautioned Zuckerberg not to "accidentally create Skynet," the military supercomputer that turns against human beings in the *Terminator* movies. "I think we can build A.I. so it works for us and helps us," Zuckerberg replied. And clearly throwing shade at Musk, he continued: "Some people fear-monger about how A.I. is a huge danger, but that seems far-fetched to me and much less likely than disasters due to widespread disease, violence, etc." Or, as he described his philosophy at a Facebook developers' conference last April, in a clear rejection of warnings from Musk and others he believes to be alarmists: "Choose hope over fear."

In the November issue of *Wired*, guest-edited by Barack Obama, Zuckerberg wrote that there is little basis beyond science fiction to worry about doomsday scenarios: "If we slow down progress in deference to unfounded concerns, we stand in the way of real gains." He compared A.I. jitters to early fears about airplanes, noting, "We didn't rush to put rules in place about how airplanes should work before we figured out how they'd fly in the first place."

Zuckerberg introduced his A.I. butler, Jarvis, right before Christmas. With the soothing voice of Morgan Freeman, it was able to help with music, lights, and even making toast. I asked the real-life Iron Man, Musk, about Zuckerberg's Jarvis, when it was in its earliest stages. "I wouldn't call it A.I. to have your household functions automated," Musk said. "It's really not A.I. to turn the lights on, set the temperature."

Zuckerberg can be just as dismissive. Asked in Germany whether Musk's apocalyptic forebodings were "hysterical" or "valid," Zuckerberg replied "hysterical." And when Musk's

SpaceX rocket blew up on the launch pad in September, destroying a satellite Facebook was leasing, Zuckerberg coldly posted that he was “deeply disappointed.”

IV. A Rupture in History

Musk and others who have raised a warning flag on A.I. have sometimes been treated like drama queens. In January 2016, Musk won the annual Luddite Award, bestowed by a Washington tech-policy think tank. Still, he’s got some pretty good wingmen. Stephen Hawking told the BBC, “I think the development of full artificial intelligence could spell the end of the human race.” Bill Gates told Charlie Rose that A.I. was potentially more dangerous than a nuclear catastrophe. Nick Bostrom, a 43-year-old Oxford philosophy professor, warned in his 2014 book, *Superintelligence*, that “once unfriendly superintelligence exists, it would prevent us from replacing it or changing its preferences. Our fate would be sealed.” And, last year, Henry Kissinger jumped on the peril bandwagon, holding a confidential meeting with top A.I. experts at the Brook, a private club in Manhattan, to discuss his concern over how smart robots could cause a rupture in history and unravel the way civilization works.

In January 2015, Musk, Bostrom, and a Who’s Who of A.I., representing both sides of the split, assembled in Puerto Rico for a conference hosted by Max Tegmark, a 49-year-old physics professor at M.I.T. who runs the Future of Life Institute, in Boston.

“Do you own a house?” Tegmark asked me. “Do you own fire insurance? The consensus in Puerto Rico was that we needed fire insurance. When we got fire and messed up with it, we invented the fire extinguisher. When we got cars and messed up, we invented the seat belt, air bag, and traffic light. But with nuclear weapons and A.I., we don’t want to learn from our mistakes. We want to plan ahead.” (Musk reminded Tegmark that a precaution as sensible as seat belts had provoked fierce opposition from the automobile industry.)

Musk, who has kick-started the funding of research into avoiding A.I.’s pitfalls, said he would give the Future of Life Institute “10 million reasons” to pursue the subject, donating \$10 million. Tegmark promptly gave \$1.5 million to Bostrom’s group in Oxford, the Future of Humanity Institute. Explaining at the time why it was crucial to be “proactive and not reactive,” Musk said it was certainly possible to “construct scenarios where the recovery of human civilization does not occur.”

Six months after the Puerto Rico conference, Musk, Hawking, Demis Hassabis, Apple co-founder Steve Wozniak, and Stuart Russell, a computer-science professor at Berkeley who co-authored the standard textbook on artificial intelligence, along with 1,000 other prominent figures, signed a letter calling for a ban on offensive autonomous weapons. “In 50 years, this 18-month period we’re in now will be seen as being crucial for the future of the A.I. community,” Russell told me. “It’s when the A.I. community finally woke up and took itself seriously and thought about what to do to make the future better.” Last September, the country’s biggest tech companies created the Partnership on Artificial Intelligence to explore the full range of issues arising from A.I., including the ethical ones. (Musk’s OpenAI quickly joined this effort.) Meanwhile, the European Union has been looking into legal issues arising from the advent of

robots and A.I.—such as whether robots have “personhood” or (as one *Financial Times* contributor wondered) should be considered more like slaves in Roman law.

At Tegmark’s second A.I. safety conference, last January at the Asilomar center, in California—chosen because that’s where scientists gathered back in 1975 and agreed to limit genetic experimentation—the topic was not so contentious. Larry Page, who was not at the Puerto Rico conference, was at Asilomar, and Musk noted that their “conversation was no longer heated.”

But while it may have been “a coming-out party for A.I. safety,” as one attendee put it—part of “a sea change” in the last year or so, as Musk says—there’s still a long way to go. “There’s no question that the top technologists in Silicon Valley now take A.I. far more seriously—that they do acknowledge it as a risk,” he observes. “I’m not sure that they yet appreciate the significance of the risk.”

Steve Wozniak has wondered publicly whether he is destined to be a family pet for robot overlords. “We started feeding our dog filet,” he told me about his own pet, over lunch with his wife, Janet, at the Original Hick’ry Pit, in Walnut Creek. “Once you start thinking you could be one, that’s how you want them treated.”

He has developed a policy of appeasement toward robots and any A.I. masters. “Why do we want to set ourselves up as the enemy when they might overpower us someday?” he said. “It should be a joint partnership. All we can do is seed them with a strong culture where they see humans as their friends.”

When I went to Peter Thiel’s elegant San Francisco office, dominated by two giant chessboards, Thiel, one of the original donors to OpenAI and a committed contrarian, said he worried that Musk’s resistance could actually be accelerating A.I. research because his end-of-the-world warnings are increasing interest in the field.

“Full-on A.I. is on the order of magnitude of extraterrestrials landing,” Thiel said. “There are some very deeply tricky questions around this If you really push on how do we make A.I. safe, I don’t think people have any clue. We don’t even know what A.I. is. It’s very hard to know how it would be controllable.”

He went on: “There’s some sense in which the A.I. question encapsulates all of people’s hopes and fears about the computer age. I think people’s intuitions do just really break down when they’re pushed to these limits because we’ve never dealt with entities that are smarter than humans on this planet.”

V. The Urge to Merge

Trying to puzzle out who is right on A.I., I drove to San Mateo to meet Ray Kurzweil for coffee at the restaurant Three. Kurzweil is the author of *The Singularity Is Near*, a Utopian vision of what an A.I. future holds. (When I mentioned to Andrew Ng that I was going to be talking to Kurzweil, he rolled his eyes. “Whenever I read Kurzweil’s *Singularity*, my eyes just naturally do that,” he said.) Kurzweil arrived with a Whole Foods bag for me, brimming with his books and

two documentaries about him. He was wearing khakis, a green-and-red plaid shirt, and several rings, including one—made with a 3-D printer—that has an *S* for his Singularity University.

Computers are already “doing many attributes of thinking,” Kurzweil told me. “Just a few years ago, A.I. couldn’t even tell the difference between a dog and cat. Now it can.” Kurzweil has a keen interest in cats and keeps a collection of 300 cat figurines in his Northern California home. At the restaurant, he asked for almond milk but couldn’t get any. The 69-year-old eats strange health concoctions and takes 90 pills a day, eager to achieve immortality—or “indefinite extensions to the existence of our mind file”—which means merging with machines. He has such an urge to merge that he sometimes uses the word “we” when talking about super-intelligent future beings—a far cry from Musk’s more ominous “they.”

I mentioned that Musk had told me he was bewildered that Kurzweil doesn’t seem to have “even 1 percent doubt” about the hazards of our “mind children,” as robotics expert Hans Moravec calls them.

“That’s just not true. I’m the one who articulated the dangers,” Kurzweil said. “The promise and peril are deeply intertwined,” he continued. “Fire kept us warm and cooked our food and also burned down our houses . . . Furthermore, there are strategies to control the peril, as there have been with biotechnology guidelines.” He summarized the three stages of the human response to new technology as Wow!, Uh-Oh, and What Other Choice Do We Have but to Move Forward? “The list of things humans can do better than computers is getting smaller and smaller,” he said. “But we create these tools to extend our long reach.”

Just as, two hundred million years ago, mammalian brains developed a neocortex that eventually enabled humans to “invent language and science and art and technology,” by the 2030s, Kurzweil predicts, we will be cyborgs, with nanobots the size of blood cells connecting us to synthetic neocortices in the cloud, giving us access to virtual reality and augmented reality from within our own nervous systems. “We will be funnier; we will be more musical; we will increase our wisdom,” he said, ultimately, as I understand it, producing a herd of Beethovens and Einsteins. Nanobots in our veins and arteries will cure diseases and heal our bodies from the inside.

He allows that Musk’s *bête noire* could come true. He notes that our A.I. progeny “may be friendly and may not be” and that “if it’s not friendly, we may have to fight it.” And perhaps the only way to fight it would be “to get an A.I. on your side that’s even smarter.”

Kurzweil told me he was surprised that Stuart Russell had “jumped on the peril bandwagon,” so I reached out to Russell and met with him in his seventh-floor office in Berkeley. The 54-year-old British-American expert on A.I. told me that his thinking had evolved and that he now “violently” disagrees with Kurzweil and others who feel that ceding the planet to super-intelligent A.I. is just fine.

Russell doesn’t give a fig whether A.I. might enable more Einsteins and Beethovens. One more Ludwig doesn’t balance the risk of destroying humanity. “As if somehow intelligence was the thing that mattered and not the quality of human experience,” he said, with exasperation. “I think

if we replaced ourselves with machines that as far as we know would have no conscious existence, no matter how many amazing things they invented, I think that would be the biggest possible tragedy.” Nick Bostrom has called the idea of a society of technological awesomeness with no human beings a “Disneyland without children.”

“There are people who believe that if the machines are more intelligent than we are, then they should just have the planet and we should go away,” Russell said. “Then there are people who say, ‘Well, we’ll upload ourselves into the machines, so we’ll still have consciousness but we’ll be machines.’ Which I would find, well, completely implausible.”

From the V.F. Summit: Elon Musk on Thinking for the Future

Russell took exception to the views of Yann LeCun, who developed the forerunner of the convolutional neural nets used by AlphaGo and is Facebook’s director of A.I. research. LeCun told the BBC that there would be no *Ex Machina* or *Terminator* scenarios, because robots would not be built with human drives—hunger, power, reproduction, self-preservation. “Yann LeCun keeps saying that there’s no reason why machines would have any self-preservation instinct,” Russell said. “And it’s simply and mathematically false. I mean, it’s so obvious that a machine will have self-preservation even if you don’t program it in because if you say, ‘Fetch the coffee,’ it can’t fetch the coffee if it’s dead. So if you give it any goal whatsoever, it has a reason to preserve its own existence to achieve that goal. And if you threaten it on your way to getting coffee, it’s going to kill you because any risk to the coffee has to be countered. People have explained this to LeCun in very simple terms.”

Russell debunked the two most common arguments for why we shouldn’t worry: “One is: It’ll never happen, which is like saying we are driving towards the cliff but we’re bound to run out of gas before we get there. And that doesn’t seem like a good way to manage the affairs of the human race. And the other is: Not to worry—we will just build robots that collaborate with us and we’ll be in human-robot teams. Which begs the question: If your robot doesn’t agree with your objectives, how do you form a team with it?”

Last year, Microsoft shut down its A.I. chatbot, Tay, after Twitter users—who were supposed to make “her” smarter “through casual and playful conversation,” as Microsoft put it—instead taught her how to reply with racist, misogynistic, and anti-Semitic slurs. “bush did 9/11, and Hitler would have done a better job than the monkey we have now,” Tay tweeted. “donald trump is the only hope we’ve got.” In response, Musk tweeted, “Will be interesting to see what the mean time to Hitler is for these bots. Only took Microsoft’s Tay a day.”

With Trump now president, Musk finds himself walking a fine line. His companies count on the U.S. government for business and subsidies, regardless of whether Marcus Aurelius or Caligula is in charge. Musk’s companies joined the amicus brief against Trump’s executive order regarding immigration and refugees, and Musk himself tweeted against the order. At the same time, unlike Uber’s Travis Kalanick, Musk has hung in there as a member of Trump’s Strategic and Policy Forum. “It’s very Elon,” says Ashlee Vance. “He’s going to do his own thing no matter what people grumble about.” He added that Musk can be “opportunistic” when necessary.

I asked Musk about the flak he had gotten for associating with Trump. In the photograph of tech executives with Trump, he had looked gloomy, and there was a weary tone in his voice when he talked about the subject. [In the end, he said](#), “it’s better to have voices of moderation in the room with the president. There are a lot of people, kind of the hard left, who essentially want to isolate—and not have any voice. Very unwise.”

VI. All About the Journey

Eliezer Yudkowsky is a highly regarded 37-year-old researcher who is trying to figure out whether it’s possible, in practice and not just in theory, to point A.I. in any direction, let alone a good one. I met him at a Japanese restaurant in Berkeley.

“How do you encode the goal functions of an A.I. such that it has an Off switch and it wants there to be an Off switch and it won’t try to eliminate the Off switch and it will let you press the Off switch, but it won’t jump ahead and press the Off switch itself?” he asked over an order of surf-and-turf rolls. “And if it self-modifies, will it self-modify in such a way as to keep the Off switch? We’re trying to work on that. It’s not easy.”

I babbled about the heirs of Klaatu, HAL, and Ultron taking over the Internet and getting control of our banking, transportation, and military. What about the replicants in *Blade Runner*, who conspire to kill their creator? Yudkowsky held his head in his hands, then patiently explained: “The A.I. doesn’t have to take over the whole Internet. It doesn’t need drones. It’s not dangerous because it has guns. It’s dangerous because it’s smarter than us. Suppose it can solve the science technology of predicting protein structure from DNA information. Then it just needs to send out a few e-mails to the labs that synthesize customized proteins. Soon it has its own molecular machinery, building even more sophisticated molecular machines.

“If you want a picture of A.I. gone wrong, don’t imagine marching humanoid robots with glowing red eyes. Imagine tiny invisible synthetic bacteria made of diamond, with tiny onboard computers, hiding inside your bloodstream and everyone else’s. And then, simultaneously, they release one microgram of botulinum toxin. Everyone just falls over dead.

“Only it won’t actually happen like that. It’s impossible for me to predict exactly how we’d lose, because the A.I. will be smarter than I am. When you’re building something smarter than you, you have to get it right on the first try.”

I thought back to my conversation with Musk and Altman. Don’t get sidetracked by the idea of killer robots, Musk said, noting, “The thing about A.I. is that it’s not the robot; it’s the computer algorithm in the Net. So the robot would just be an end effector, just a series of sensors and actuators. A.I. is in the Net The important thing is that if we do get some sort of runaway algorithm, then the human A.I. collective can stop the runaway algorithm. But if there’s large, centralized A.I. that decides, then there’s no stopping it.”

Altman expanded upon the scenario: “An agent that had full control of the Internet could have far more effect on the world than an agent that had full control of a sophisticated robot. Our lives

are already so dependent on the Internet that an agent that had no body whatsoever but could use the Internet really well would be far more powerful.”

Even robots with a seemingly benign task could indifferently harm us. “Let’s say you create a self-improving A.I. to pick strawberries,” Musk said, “and it gets better and better at picking strawberries and picks more and more and it is self-improving, so all it really wants to do is pick strawberries. So then it would have all the world be strawberry fields. Strawberry fields forever.” No room for human beings.

But can they ever really develop a kill switch? “I’m not sure I’d want to be the one holding the kill switch for some superpowered A.I., because you’d be the first thing it kills,” Musk replied.

Altman tried to capture the chilling grandeur of what’s at stake: “It’s a very exciting time to be alive, because in the next few decades we are either going to head toward self-destruction or toward human descendants eventually colonizing the universe.”

“Right,” Musk said, adding, “If you believe the end is the heat death of the universe, it really is all about the journey.”

The man who is so worried about extinction chuckled at his own extinction joke. As H. P. Lovecraft once wrote, “From even the greatest of horrors irony is seldom absent.”